(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 073 002 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
31.01.2001 Bulletin 2001/05

(51) Int. Cl.7: G06K 9/34

(21) Application number: 00304765.1

(22) Date of filing: 06.06.2000

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 30.07.1999 JP 21715199

(71) Applicant:
KABUSHIKI KAISHA TOSHIBA
Kawasaki-shi, Kanagawa-ken 210-8572 (JP)

(72) Inventor:
Akagi, Takuma,
IPD, Toshiba Corporation
Tokyo (JP)

(74) Representative: Shindler, Nigel
BATCHELLOR, KIRK & CO.
102-108 Clerkenwell Road
London EC1M 5SA (GB)

(54) Character extracting method

(57) In a character recognition device for the character recognition based on a light and shade image of an input character image, the light and shade image of the input character image is separated into a character area and a background area using at least density values of pixels and then, the above-mentioned character area is separated again into more than two areas using at least density values of pixels and based on an area information obtained by this re-separation, the characters are separated to individual characters.

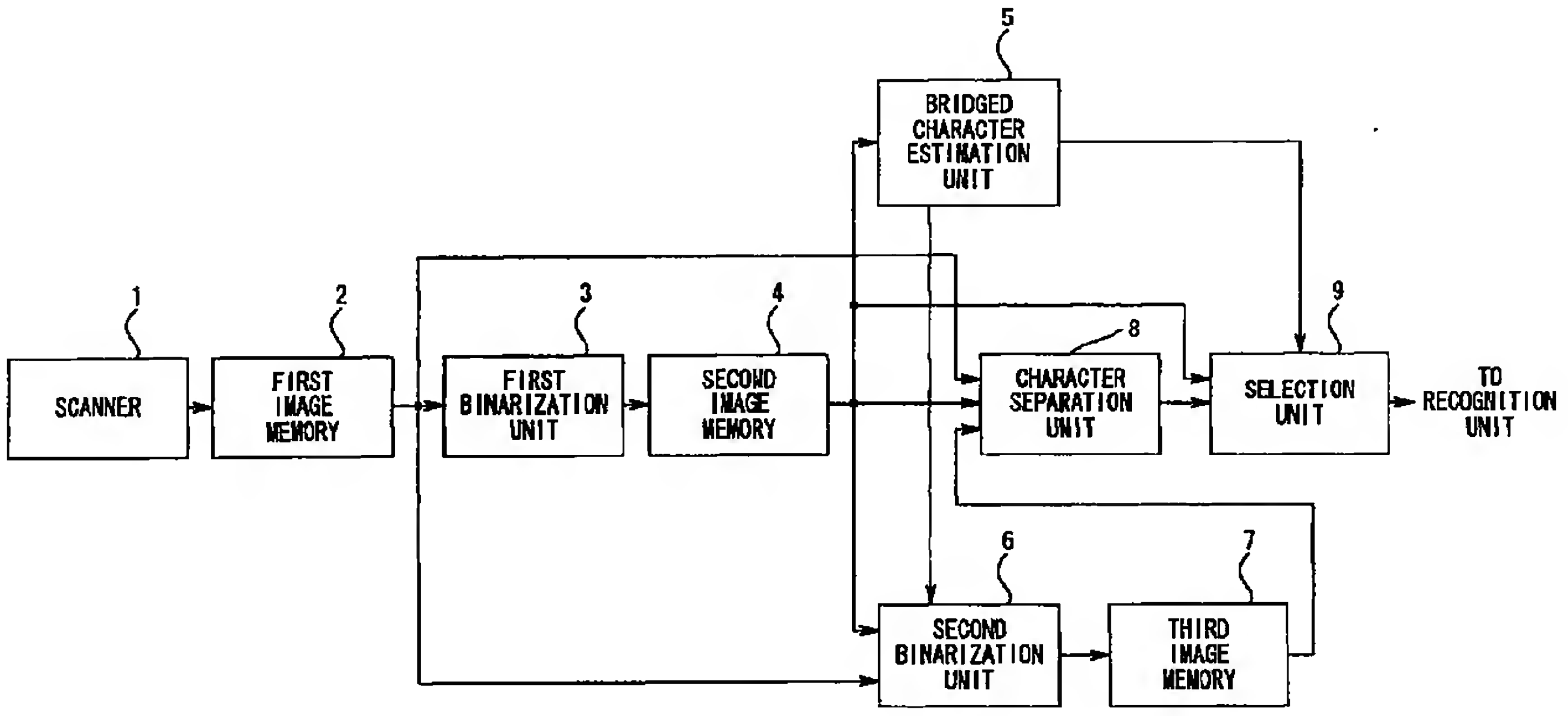

FIG. 5

Printed by Xerox (UK) Business Services
2.16.7 (HRS)/3.6

Description

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The present invention relates to a character extracting method to extract bridged characters to each character at a time in a character recognition device which recognizes characters according to, for instance, input light and shade character images.

2. Description of the Related Art

[0002] Generally, the character recognition technology is broadly divided into input of character images, extraction of character lines, character extracting and character recognition.

[0003] Generally, bridged characters are extracted after binarizing input character images. The input character images are divided into a character area and a background area. At this time, plural characters are extracted to one character area in the bridged stated. In the prior art, these bridged characters are separated each other by obtaining structural and analytical separating points according to the shape of bridged characters. That is, when the shape of bridged characters have a dent such as valley, a character area was cut out to each character by judging the dent as the junction point.

[0004] However, a variety of patterns are considered for the shapes of junction points of bridged characters and the patterns that could be separated according to the prior art are limited only to a few special shapes out of these patterns.

[0005] Generally, in the case of document images of handwritten characters, for the most part of bridged characters when extracting them, characters on an input document image are already bridged from the first time when the document is input.

[0006] On the contrary, in the case of document images in printing type, the most part of causes for bridged characters that become a problem when characters are extracted are not actually bridged but are due to a low resolution of a scanner (a character image input device) and a binarization miss during the binarization processing.

[0007] Therefore, if returning back to light and shade images that are binarized input character images, a bridged point between characters should be detectable relatively easily.

[0008] However, when light and shade images only are simply used, a problem becomes more complicate including specifying of a character area, delay of processing speed and in addition, such ill effects as an error in character position estimation, etc. are caused and the performance is rather deteriorated.

SUMMARY OF THE INVENTION

[0009] It is, therefore, an object of the present invention to provide a character extracting method that is able to highly precisely and efficiently find out bridged characters that are difficult to find from the shapes and separate them.

[0010] According to the present invention, there is provided a character extracting method in a character recognition device for recognizing characters according to input character images, comprising a first step for separating the input character iamges into a character area and a background area; a second step for separating the character area separated in the first step into more than two areas using density values of the pixels of the character images in the character area; and a third step for separating the character area into a character and a character according to the area information obtained in the second step.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011]

FIG. 1 is an exlarged view showing an example of an input document image before inputting by a scanner;

FIG. 2 is a diagram for explaining the state to input a document image by marking off with meshes;

FIG. 3 is a diagram showing an example of a light and shade image input by a scanner;

FIG. 4 is a diagram showing an example of a binarized image by binarizing the light and shade image shown in FIG. 3;

FIG. 5 is a block diagram schematically showing a structure of a character extracting device to which the character extracting method of the present invention is applicable;

FIG. 6 is a flowchart for explaining the character extracting method of the present invention;

FIG. 7 is a diagram showing an example of a first binarized image that is binarized in the first binarization unit;

FIG. 8 is a diagram showing an example of a second binarized image that is binarized in the second binarization unit; and

FIG. 9 is a flowchart for explaining a method for deciding a bridged character position in a character separation unit.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0012] Preferred embodiments of the present invention will be described below referring to the attached drawings.

[0013] Further, in the following description it is assumed that the more small density values of pixels are, the more black those pixels will become, and the more large density values are, the more white the pixels will become.

[0014] FIG. 1 shows an input document image enlarged before inputting by a scanner. Figures "1" and "2" are written closely each other; however, they are not bridged each other as there is a blank space between them. Now, let's consider a case to read such document images with a scanner.

[0015] A scanner divides document images in such meshes as shown in FIG. 2, takes a mean value of densities in respective rectangles and inputs the mean density value as a representative density of pixels corresponding to the rectangles. Shown in FIG. 3 is an example of a light and shade image input by a scanner. In this example, a light and shade image is a vertically 6-dot and laterally 12-dot image.

[0016] In FIG. 3, the portion a painted out in black represents the most large density pixels, the portion b shown by cross oblique lines represents pixels in density smaller than the pixels a, the portion c shown in thick oblique lines represents pixels in density smaller than the pixels b and the portion d shown fine oblique lines represents pixels in density smaller than the pixels c, respectively.

[0017] When the width of the blank space between characters is thinner than rectangles taken in by a scanner, pixels in small density are obtained even when they are in the blank space portion as seen in FIG. 3. Therefore, when this image is binarized, a binarized image in a shape of two bridged characters is obtained as shown in FIG. 4.

[0018] In a conventional character extracting method, bridged characters were separated using only a binarized image shown in FIG. 4.

[0019] However, when looking the light and shade image shown in FIG. 3, the contour of a character is in a density larger than that of the central portion of the character. The bridged portion of characters is also not exceptional and pixels corresponding to the blank space are in a density larger than that of the character portion.

[0020] Therefore, in the present invention, portions in large density are found efficiently and by deciding a bridged character portion, a character area is separated to each character. This will be explained below in detail.

[0021] FIG. 5 is a block diagram showing the structure of a character extracting device to which the character extracting method of the present invention is applicable. In FIG. 5, a scanner 1, as a character image input means, reads an input document image optically and inputs such a light and shade image as shown in FIG. 3. The light and shade image input by the scanner 1 is temporarily stored in a first image memory 2.

[0022] A first binarization unit 3 binarizes the light and shade image temporarily stored in the first image memrory 2 according to, for instance, a known Ohtsu's binarization method (Refer to "Automatic Threshold Selection Method according to Discrimination and Minimum Square Standard", by Nobuyuki Ohtsu, Shingakuron (D), vol. J63-D, no. 4, pp. 349-356, 1980) and outputs such a first binarized image as shown in FIG. 4.

[0023] Furthermore, the binarization method for the light and shade image is disclosed in U.S.Patent No. 5,784,500 (Jul. 21, 1988).

[0024] The first binarized image output from the first binarization unit 3 is stored in a second image memory 4 temporarily. A bridged character estimation unit 5 estimates (judges) if there are bridged characters based on the first binarized image temporarily stored in the second image memory. Regarding the standard for estimation, in the case where a lateral size of a black pixel area is wider than a vertical size, it is estimated that plural characters are bridged.

[0025] A second binarization unit 6 operates when the bridged character estimation unit 5 estimates that plural characters are bridged, and when a light and shade image temporarily stored in the first image memory 2 and the first binarized image temporarily stored in the second image memory 4 are input, density values are first taken only from light and shade image pixels corresponding to positions of black pixels of the first binarized image and registered for a histogram.

[0026] Then, based on the obtained histogram, the second binarization unit 6 obtains a threshold value to halve the density value of the histogram using the above-mentioned known Ohtsu's binarization method. Only when a density value of the same coordinates of the light and shade image out of the black pixel of the binarized image is largeer than the obtained threshold value, the black pixels are reversed to white pixels and a new binarized image (a second binarized image) is formed.

[0027] The second binarized image output from the second binarization unit 6 is stored temporarily in a third image memory 7. A character separation unit 8 decides bridged character positions according to the light and shade image temporarily stored in the first image memory 2, the first binarized image temporarily stored in the second image memory 4 and the second binarized image temporarily stored in the third image memory 7 and separates the character images to each character according to the information on this decided character bridge position. What is separated at this time is the character image (the first binarized image) obtained in the first binarization unit 3.

[0028] A selection unit 9 selects the stored contents of the second image memory 4 or the output of the character separation unit 8 based on the result of estimation by the bridged character estimation unit 5. That is, when the bridged character estimation unit 5 estimated that there are no bridged characters, the stored contents of the second image memory 4 are selected and when it is estimated that there are bridged characters, the output of the character separation unit 8 is selected.

[0029] Next, the character extracting method of the present invention will be described in detail referring to a flowchart shown in FIG. 6. First, a light and shade character image is input to the scanner 1 and stored in the first image memory 2 temporarily (S1). Then, the light and shade image in the first image memory 2 is converted into the first binarized image in the first binarization unit 3 using the Ohtsu's binarization method and stored in the second image memory 4 temporarily (S2).

[0030] Then, the bridged character estimation unit 5 judges if there are bridged characters based on the obtained first binarized image in the second image memory 4 (S3). Regarding the standard for judging bridged characters, when a lateral size of a black pixel area is wider than a vertical size, it is judged that plural characters are bridged. When judged there are no bridged characters, the processing is terminated. In this case, the selection unit 9 selects the first binarized image in the second image memory 4 and forward it to the next recognition processing.

[0031] When judged there are bridged characters, the second binarization unit 6 takes density values only from light and shade image pixels corresponding to the positions of the black pixels of the first binarized image obtained in the first binarization unit 3 and registers for a histogram (S4). Then, based on the obtained histogram, a threshold value to halve the density value of the histogram is obtained using the Ohtsu's binarization method. When the density values of the same coordinates of the light and shade image out of the black pixels of the binarized image are larger than the threshold value obtained, the black pixels are reversed to the white pixels, and the second binarized image is formed and stored in the third image memory 7 temporarily (S5).

[0032] FIG. 7 shows an example of the first binarized image binarized in the first binarization unit 3 and FIG. 8 shows an example of the second binarized image binarized in the second binarization unit 6, and in FIGs. 7 and 8, the black rectangles are portions that are regarded to be black pixels in the binarization and white rectangles are portions that are regarded to be white pixels in the binarization.

[0033] When a density histogram is plotted again for a character area only and binarized, the portions that were made black pixels by the quantization error of the scanner as mentioned above; that is, the closed portions between characters or the contours of the characters are changed into the white pixels because of a density value. Therefore, in FIG. 8, it can be seen that a new blank space is produced between characters "2" and "5" and "5" and "3" (the arrow portions in FIG.).

[0034] Then, in the character separation unit 8, a bridged character position is decided (S6) based on the second binarized image obtained in the second binarization unit 6, the first binarized image obtained in the first binarization unit 3 and the light and shade image input with the scanner 1. The decision of bridged character position will be explained later in detail. Then, based on the information on the decided bridged character position, character images (the first binarized image obtained in the first binarization unit 3) is separated to each character (S7). In this case, the selection unit 9 selects the output of the character separation unit 8 and forward it to the next recognition processing.

[0035] Next, the bridged character position deciding method in the character separation unit 8 will be explained in detail referring to a flowchart shown in FIG. 9. First, the first binarized image obtained in the first binarization unit 3 is compared with the second binarized image obtained in the second binarization unit 6 and columns containing much pixels that are newly turned to white pixels are detected (S11). Then, the columns detected in Step S11 are checked if there are single black pixels in the vertical direction of the second binarized image (S12).

[0036] When no single black pixels are detected in the vertical direction as a result of the above check, the operation proceeds to Step S14. When black pixels are detected, the light and shade image input by the scanner 1 is checked (S13). That is, a mean density value at the positions of black pixels in the first binarized image of said column at the same column of the light and shade image is obtained. The same processing is executed for several columns that are left and right to said column. When said column is a ridge to the left and right columns; that is, it is judged whether a mean value of the density values of the left and right columns is smaller than a mean value of density values of said columns. When the mean value is smaller as a result of this judgement, proceeds to Step S14 and when not smaller, the image is withdrawn from a candidate for separation (the characters are not separated).

[0037] Then, whether there are characters at the left and right side (both sides) of the column obtained from the process in Step S12 or Step S13 is checked (S14). The white pixels obtained from the second binarized image are presented at the character edge in addition to the bridged point of characters as shown in FIG. 8. It is therefore necessary to check if there are characters at both ends of said column. For instance, the number of black pixels are counted over several columns at both ends of said column based on the first binarized image and if there were black pixels more than a certain value, it is regarded that there are characters.

[0038] When no character is found as a result of the check in Step S14, the image is withdrawn from a candidate for separation (no character is separated) and when there are characters, the character separation process is executed by the character separation unit (S15).

[0039] Further, if the resolution of the scanner 1 that inputs character images is small in Step S12 of the flowchart shown in FIG. 9, jump to Step S14 even when there are some black pixels. In this case, as a standard for deciding a separation candidate, a difference between the black pixels of the first binarized image obtained in the first binarization unit 3 and the black pixels of the second binarized image obtained in the second binarization unit 6 is taken and a column having a larger difference is made a candidate column for the character separation. In this case, a threshold value showing the size of difference is lowered so that the resolution of the scanner 1 is low.

[0040] As described above, according to the above-mentioned embodiment, it is possible to separate bridged characters that could not be separated so far by a binarized image only. Further, the bridged character separation can be processed with considerably lighter load than all processes using a light and shade image only. Accordingly, the bridged characters that are difficult to find from the viewpoint of shape can be found and separated highly precisely and efficiently.

[0041] As described above in detail, according to the present invention, a character extracting method capable of finding and separating bridged characters highly precisely and efficiently from the viewpoint of shape can be provided.

Claims

1. A character extracting method in a character recognition device for recognizing characters according to input character images, comprising:

   a first step for separating the input character iamges into a character area and a background area;
   a second step for separating the character area separated in the first step into more than two areas using density values of the pixels of the character images in the character area; and
   a third step for separating the character area into a character and a character according to the area information obtained in the second step.

2. A method according to claim 1, wherein the third step includes a step of separating the characters into individual chracters using a pixel line or a pixel column as a separating point, which has a large difference between the character area obtained in the first step and the area obtained in the second step.

3. A method according to claim 1, wherein the third step includes a step of comparing between density values at both side of a region to be separated and a density value of the region itself to be separated and separating the characters into individual characters only when the density values at both sides are smaller than the density value of the region itself to be separated.

4. A method according to claim 1, wherein the second step includes a step of separating the characters into individual chracters based on a histogram which is formed using density values of the pixels at the positions applicable to the character area.

5. A method according to claim 2, wherein when deciding a separating point based on a difference between the density value of the character area obtained in the first step and the density value of the area obtained in the seocnd step, a threshold value for deciding a separating point is varied based on a resolution value of an input character image.

6. A character extracting device in a character recognition device for recognizing characters according to input character images, comprising:

   first means for separating the input character iamges into a character area and a background area;
   second means for separating the character area separated by the first means into more than two areas using density values of the pixels of the character images in the character area; and
   third means for separating the character area into a character and a character according to the area information obtained in the second means.

7. A device according to claim 6, wherein the third means includes means for separating the characters into individual chracters using a pixel line or a pixel column as a separating point, which has a large difference between the character area obtained by the first means and the area obtained by the second means.

8. A device according to claim 6, wherein the third means includes means for comparing between density values at both side of a region to be separated and a density value of the region itself to be separated and separating the characters into individual characters only when the density values at both sides are smaller than the density value of the region itself to be separated.

9. A device according to claim 6, wherein the second means includes means for separating the characters into individual chracters based on a histogram which is formed using density values of the pixels at the positions applicable to the character area.

10. A device according to claim 7, wherein when deciding a separating point based on a difference between the density value of the character area obtained by the first means and the density value of

the area obtained by the seocnd means, a threshold value for deciding a separating point is varied based on a resolution value of an input character image.

11. A character separating method comprising the steps of:

storing an input character images in a first memory;
converting the character iamges stored in the first memory into first binarized images and storing the first binarized images in a second memory;
judging whether character images are bridged based on the first binary images stored in the second memory;
prepareing a histogram by obtaining density values from pixels of character images equivalent to the positions of black pixels of the first binarized images stored in the second memory when the characters are judged as being bridged in the judging step;
obtaining a threshold value for halving the density values on the histogram based on the prepared histogram;
preparing a second binarized image by reversing black pixels to white pixels and storing the second binarized image in a third memory only when density values at the same coordinates of the character images out of the black pixel of the binarized image is largeer than the obtained threshold value;
deciding the position of bridged characters according to on the second binarized image, the first binarized image and the character images stored in the first memory; and
separating the first binarized image into individual character images according to the information on the decided bridged character position.

12. A method according to claim 11, wherein the step of deciding the position of bridged characters includes:

comparing the first and second binarized images to detect a column containing newly reversed white pixels much;
checking the detected column to determine if there is even a single black pixel in the vertical direction of the second binarized image;
judging the character images stored in the first memory as to whether a mean value of density values of the left and right columns is lower than a mean value of the density of respective columns when the black pixels are detected as a result of the check;
counting the number of black pixels over several columns at both ends of the respective columns based on the first binarized image; and
regarding that there exist characters at both ends of the column if there are black pixels more than a certain values as the result of the step of counting the number of black pixels.

13. A character extracting device in a character recognizing device comprising:

a scanner to obtain character images by optically reading input document images;
a first image memory to store the character images obtained by the scanner;
a first binarization unit to binarize the character images stored in the first image memory and output a first binarized image;
a second image memory to store the first binarized image output from the first binarization unit;
a bridged characters estimation unit to estimate whether there are bridged characters according to the first binarized image stored in the second image memory;
a second binarization unit to prepare a second binarized image by reversing the black pixels to the white pixels when the bridged character estimation unit estimated bridged plural characters, density values are taken only from the pixel of the character images at a position equivalent to that of the black pixel of the first binarized image based on the character images stored in the first image memory and the first binarized image stored in the second image memory and a histogram is registered, and based on the registered histogram, out of the black pixels of the first binarized image, a density value of the same coordinates of the character images is higher than the obtained threshold value;
a third image memory to store the second binarized image prepared by the second binarization unit; and
a character separation unit to decide a bridged character position based on the character images stored in the first image memory, the first binarized image stored in the second image memory and the second binarized image stored in the third image memory and separate bridged character image into individual characters according to the decided character position information.

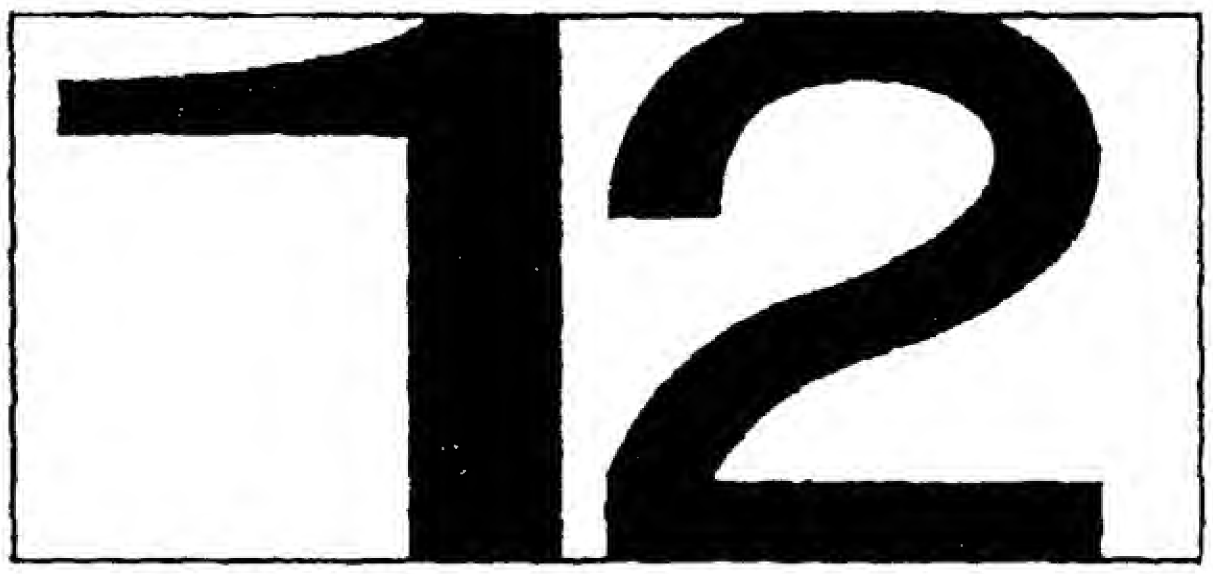

FIG. 1

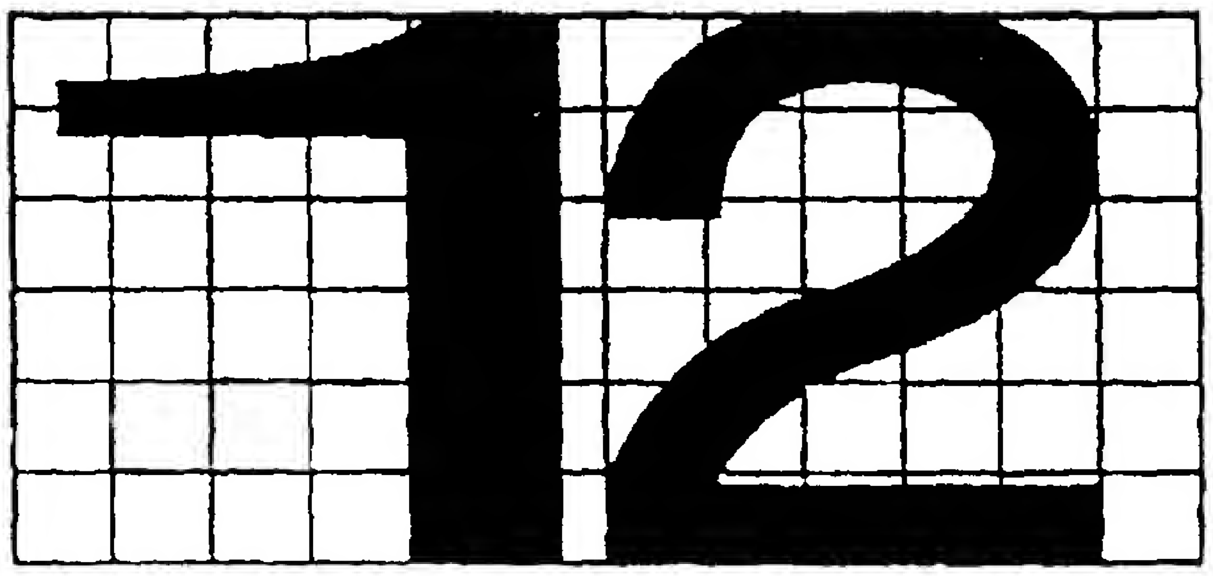

FIG. 2

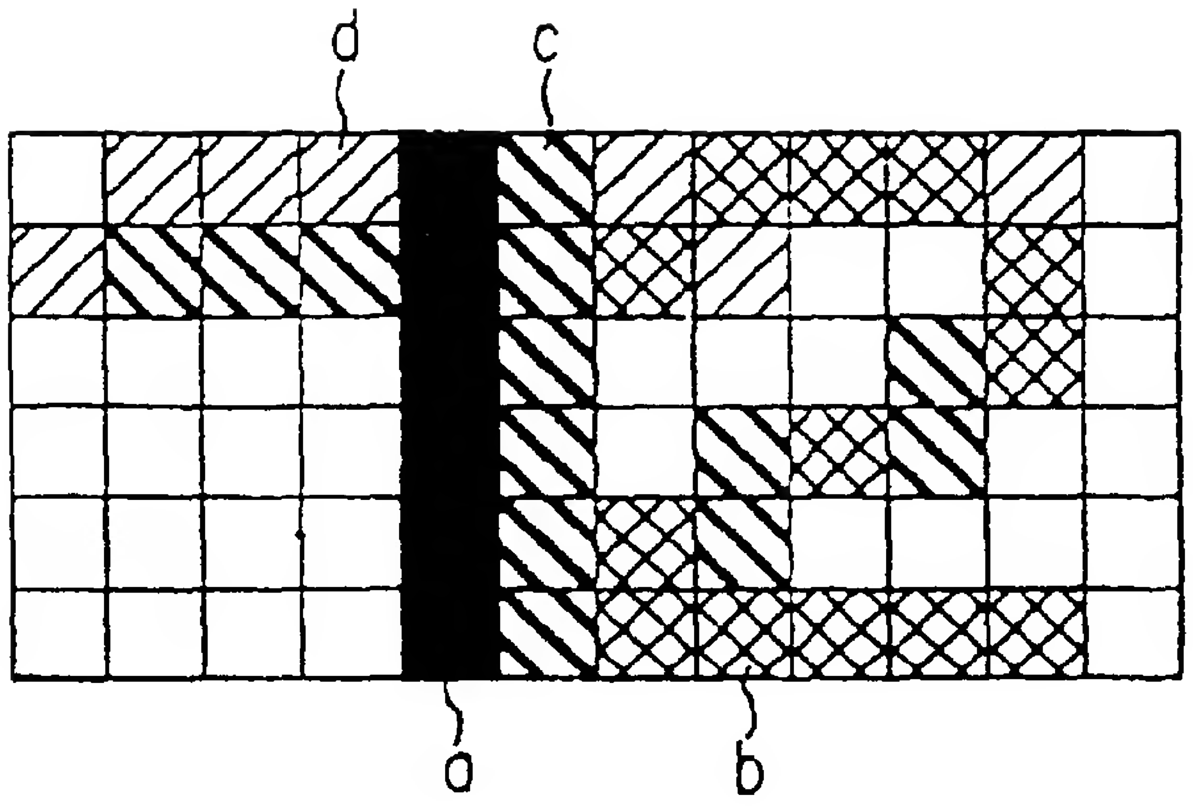

FIG. 3

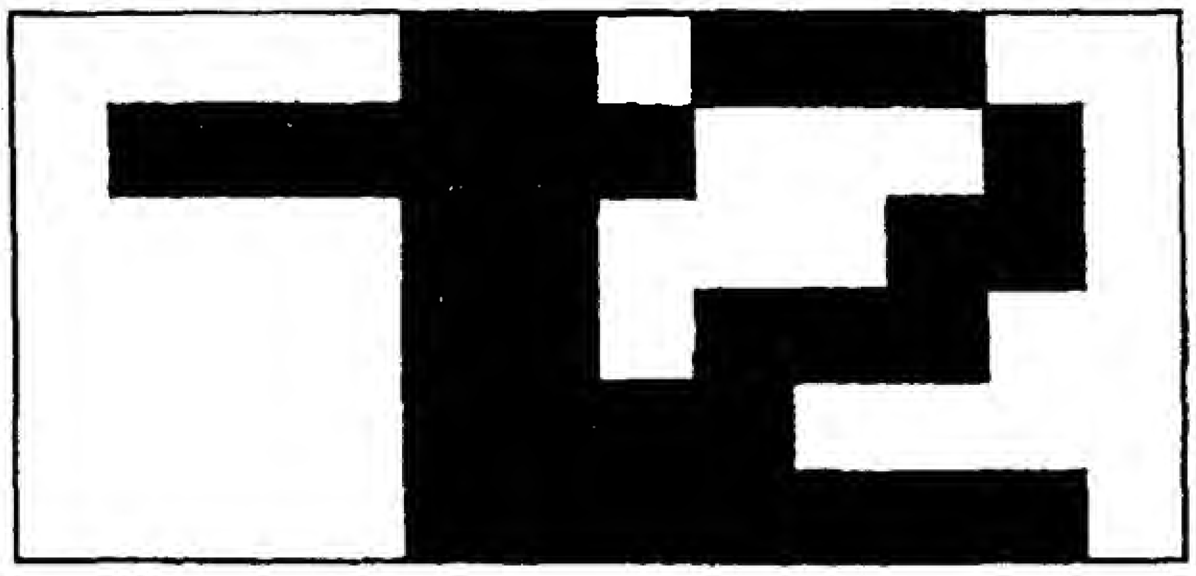

FIG. 4

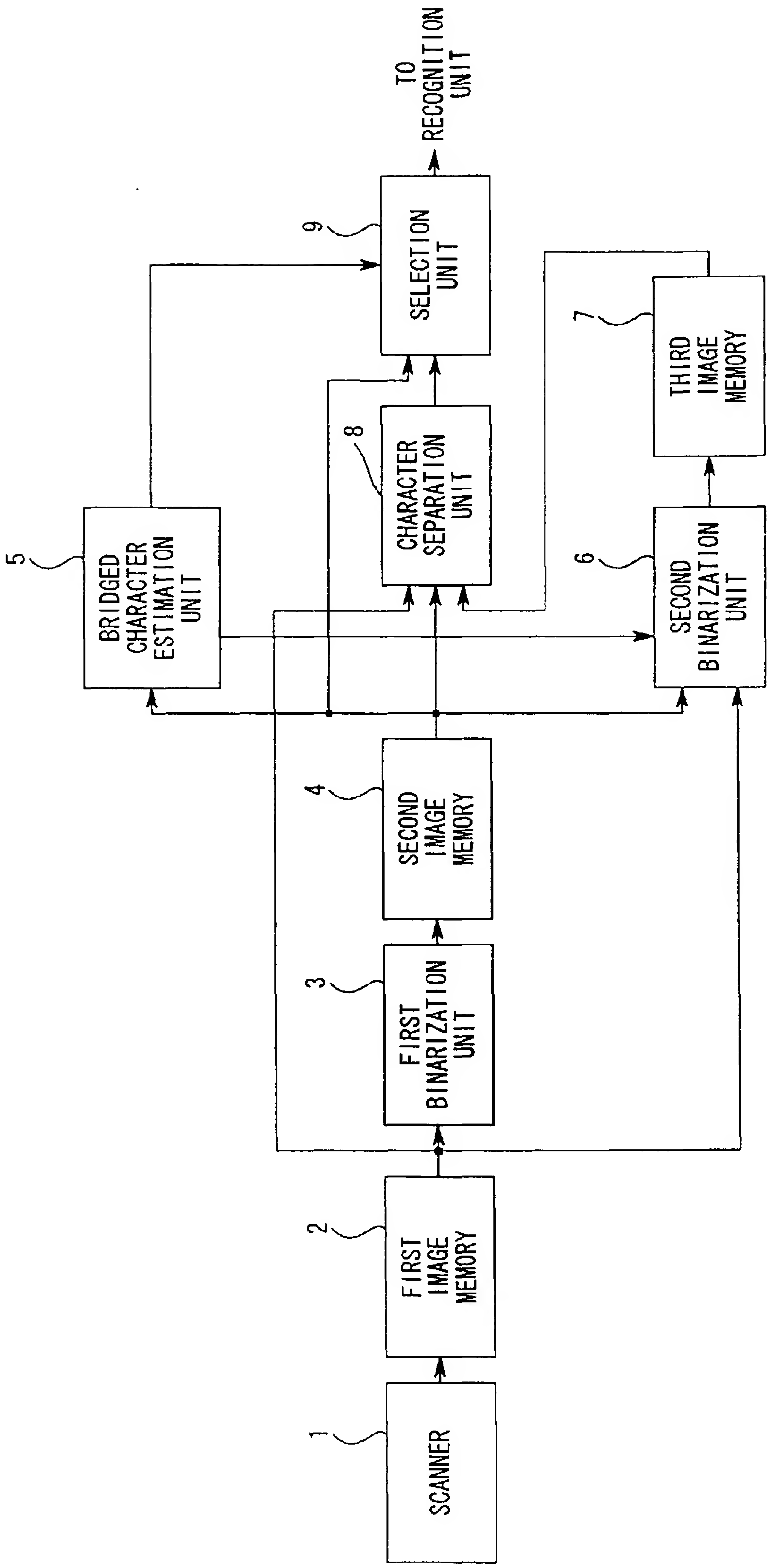
1
SCANNER
2
FIRST IMAGE MEMORY
3
FIRST BINARIZATION UNIT
4
SECOND IMAGE MEMORY
5
BRIDGED CHARACTER ESTIMATION UNIT
6
SECOND BINARIZATION UNIT
7
THIRD IMAGE MEMORY
8
CHARACTER SEPARATION UNIT
9
SELECTION UNIT
TO RECOGNITION UNIT

FIG. 5

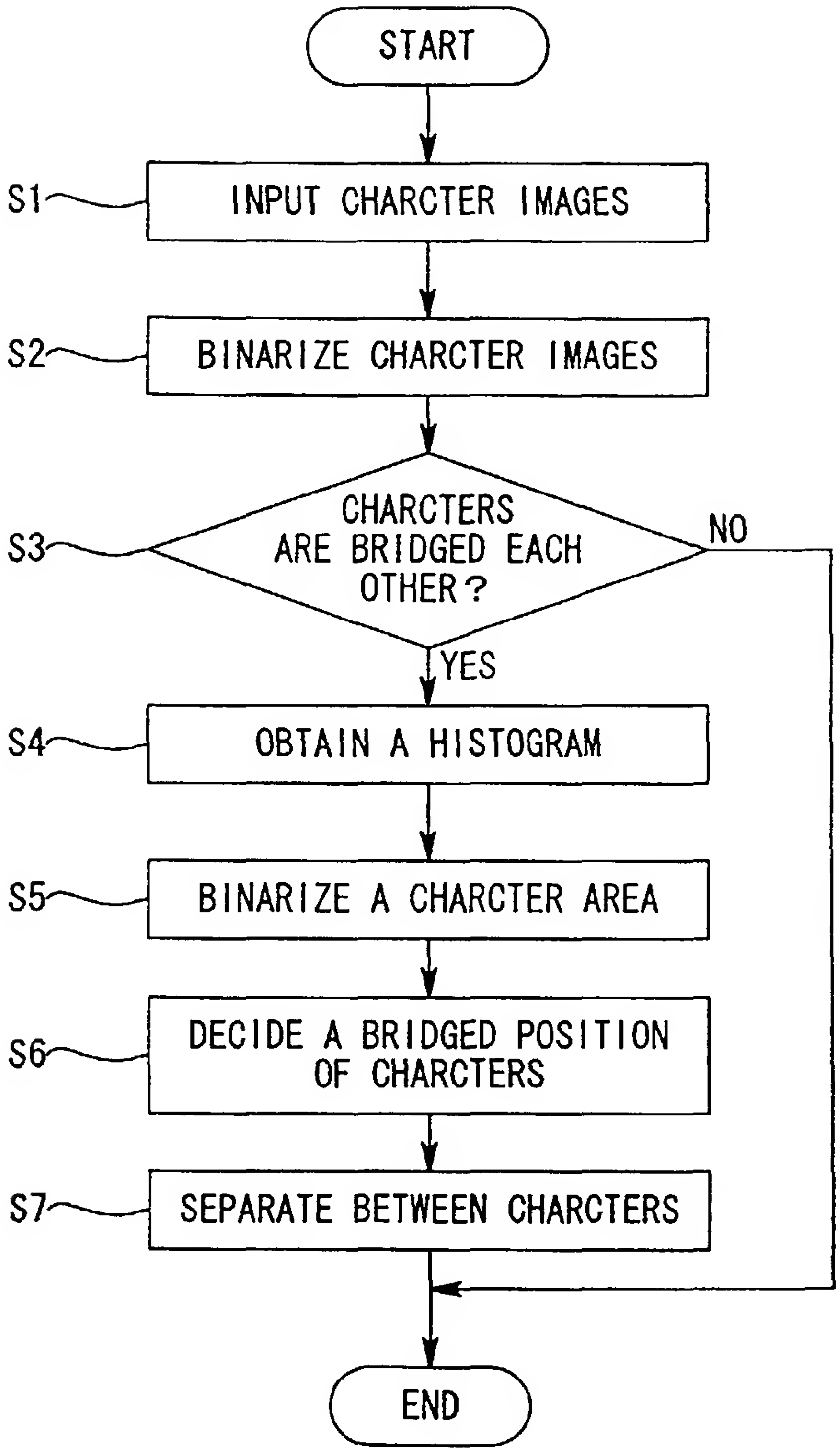
START
S1
INPUT CHARCTER IMAGES
S2
BINARIZE CHARCTER IMAGES
S3
CHARCTERS ARE BRIDGED EACH OTHER?
NO
YES
S4
OBTAIN A HISTOGRAM
S5
BINARIZE A CHARCTER AREA
S6
DECIDE A BRIDGED POSITION OF CHARCTERS
S7
SEPARATE BETWEEN CHARCTERS
END

FIG. 6

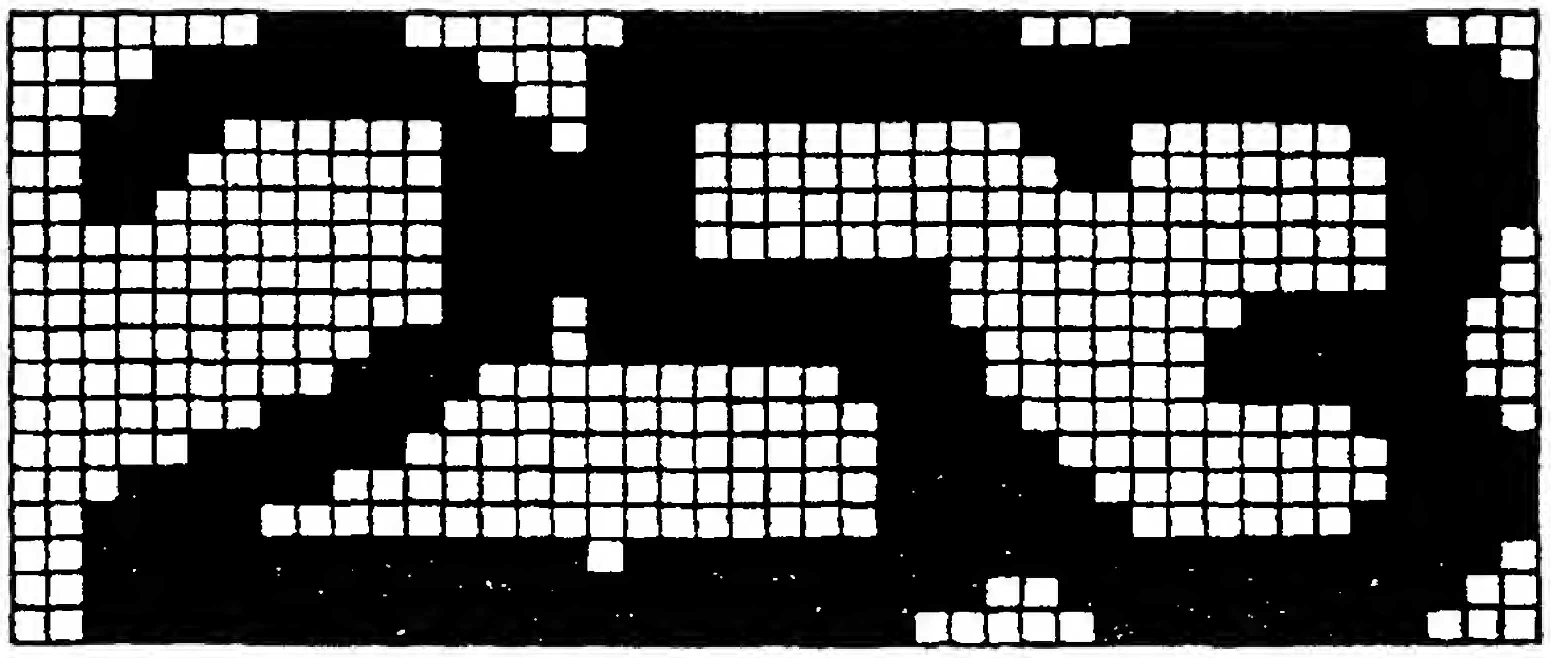

FIG. 7

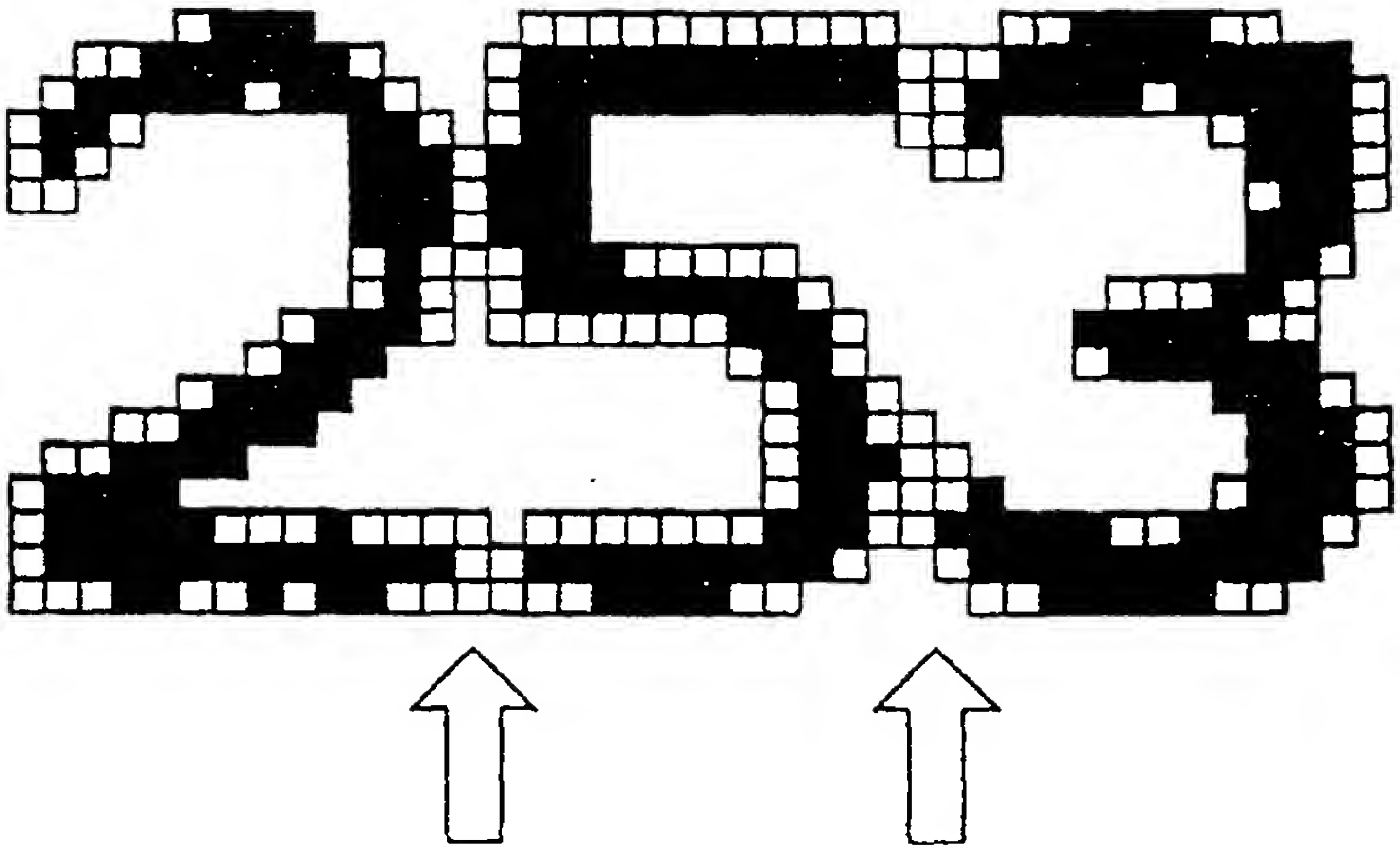

FIG. 8

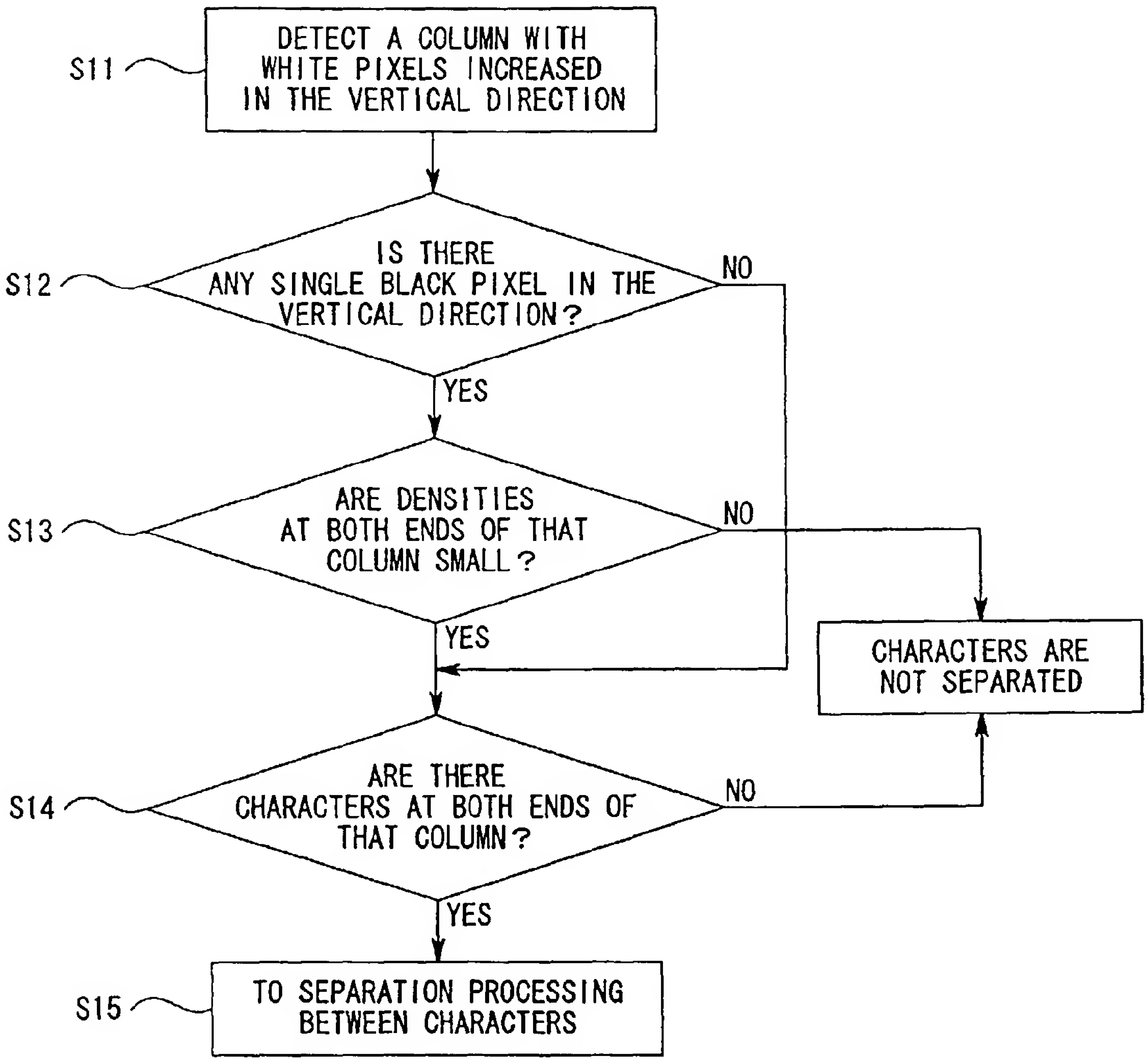
S11
DETECT A COLUMN WITH WHITE PIXELS INCREASED IN THE VERTICAL DIRECTION
S12
IS THERE ANY SINGLE BLACK PIXEL IN THE VERTICAL DIRECTION?
NO
YES
S13
ARE DENSITIES AT BOTH ENDS OF THAT COLUMN SMALL?
NO
YES
CHARACTERS ARE NOT SEPARATED
S14
ARE THERE CHARACTERS AT BOTH ENDS OF THAT COLUMN?
NO
YES
S15
TO SEPARATION PROCESSING BETWEEN CHARACTERS

FIG. 9